

NAVAL HEALTH RESEARCH CENTER

WALK TESTS AS INDICATORS OF AEROBIC CAPACITY

R. R. Vickers. Jr.

20040319 068

Report No. 02-22

Approved for public release; distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (MED-02)
2300 E ST. NW
WASHINGTON, DC 20372-5300



WALK TESTS AS INDICATORS OF AEROBIC CAPACITY

Ross R. Vickers, Jr.
Human Performance Department
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122
e-mail: **Vickers@nhrc.navy.mil**

Report No. 02-22, supported by the U.S. Marine Corps under research work unit 60109. The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

Approved for public release; distribution unlimited

SUMMARY

Background

Measures of cardiorespiratory fitness are routinely included in physical fitness tests (PFTs) that are administered for personnel selection or to monitor the fitness of a population. Typically, the cardiorespiratory measures take the form of a run test. Walk tests may be a viable alternative to run tests. However, much of the literature on walk tests is recent. To date, walk test validity has not been directly compared with run test validity.

Objective

This report provides a quantitative summary of the validity of walk tests and compares walk test validity with run test validity.

Approach

The published literature was reviewed to identify studies that related walk test performance to laboratory measures of maximal oxygen uptake capacity (VO_{2max}). Meta-analysis techniques were used to average the reported correlation coefficients and compare them with the average values of the same statistics for run tests.

Findings

The literature search produced 39 studies, 37 of which concerned 1-km, 2-km, 1-Mile, 6-min, or 12-min walk tests. Walk test performance was significantly ($p < 10^{-6}$) related to VO_{2max} for each of those tests. The relationships were near the lower boundary (i.e., $r = .60$) for acceptable validity. Each walk test was less valid than its comparable run test. However, combining walk test performance with age, weight, gender, and exercise heart rate produced regression equations that predicted VO_{2max} as well as run tests. Standard errors of estimate were $5.01 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ for the walk test for men and $3.78 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ for women. The comparable run test values were $4.69 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ and $3.38 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$, respectively.

Conclusions

Walk tests are valid indicators of maximal aerobic capacity. However, walk test performance must be combined with information on age, weight, gender, and exercise heart rate to produce VO_{2max} estimates that are as good as the run tests currently used in PFTs. The multivariate approach would be recommended when using walk tests.

Introduction

Running performance is commonly used to assess aerobic fitness in military physical fitness tests (PFTs). A substantial body of evidence relating running performance to measured maximal oxygen uptake capacity (VO_{2max}) supports this practice. Walk tests are an alternative method of estimating aerobic fitness that may be preferable in some situations. Solway, Brooks, et al. (2001) provided a qualitative review of the evidence supporting the claim that walk tests are valid indicators of VO_{2max} . This review provides a quantitative summary of that evidence and a comparison of walk tests and run tests.

This report focuses on walk test validity. In everyday conversation, the word "valid" conveys the idea that an assertion is "true," or "correct." Valid has a narrower technical definition when used in connection with testing standards. In this context, "validity" refers to the appropriateness of some interpretation of a set of test scores (American Psychological Association, 1985). Test validation is the process of gathering empirical evidence to support the proposed interpretation(s) of the scores.

Good testing practice requires that the validity of walk tests be demonstrated empirically. Evidence that walk test performance is reliably related to laboratory VO_{2max} measures a critical requirement for establishing the validity of walk tests. This evidence is critical because laboratory measurements of oxygen uptake during treadmill runs or bicycle ergometer rides are accepted as the best available methods of assessing aerobic fitness. Walk tests would not be plausible indicators of aerobic fitness if performance were not related to this accepted standard. Therefore, this report uses meta-analytic procedures (cf., Cooper & Hedges, 1994; Hedges & Olkins, 1985) to summarize the available evidence bearing on the claim that walk tests meet this basic validity requirement.

Walk Test Validity Evidence

Any review begins with a search for relevant studies. For the present purposes, a relevant study was one that reported an empirical estimate of the association between walk test performance and VO_{2max} . An initial list of relevant studies was constructed from the Solway, Brooks, et al. (2001) reference list. This list was extended by searching the PubMed® database using "walk test" as the search term. The abstracts for the articles identified in this search were examined to determine whether VO_{2max} had been measured. If so, the study was added to the list.

Copies of the articles in the original list were obtained. The articles were read to determine which ones reported the required correlations. When a correlation was reported, the paper was read to identify any references to prior studies of the performance-VO_{2max} relationship. Citations not previously identified were added to the list of studies to be reviewed.

The list of relevant articles was completed by a further search of the PubMed database. PubMed includes a "related articles" function. Once an article of interest has been identified, clicking a button generates a list of other articles dealing with similar subject matter. This function was used for each study identified in the PubMed search. If the abstract of a related article suggested that a relevant correlation might be reported, the article was examined to determine whether it provided evidence that should be added to the database for this review.

The search identified 39 studies that reported at least 1 correlation between walk test performance and VO_{2max} (Appendix A). The cumulative sample size was 1,927 participants. The samples were not representative of the general population. Most ($n = 1,117$, 58.0%) participants were patients with moderate to severe cardiac or respiratory disease. The average age of the participants ranged from 7 years to 68 years, but most data were from samples near the extremes of this range (<15 years, $n = 221$, 11.5%; >50 years, $n = 995$, 51.6%). Adult samples with average ages between 36 and 50 ($n = 628$, 32.6%) accounted for most of the remaining data. Only about 1 of every 25 ($n = 83$, 4.3%) participants was from a sample of young adults. Because patient populations tended to be older, the typical study participant was a patient over the age of 50.

Table 1 presents the basic validity evidence. Table 2 summarizes that evidence on a test-by-test basis. The cumulative evidence leaves no doubt that walk tests are related to VO_{2max}. Major observations were:

- A. The average validity coefficient was highly significant ($p < 10^{-6}$)¹ for each test that has been studied in more than one sample.
- B. The average validity coefficients differed significantly between tests ($\chi^2 = 13.29$, 4 *df*, $p < .010$).²
- C. The run test was more valid than the walk test for the 1-km ($z = 3.33$, $p < .001$), 1-mile ($z = 3.88$, $p < .001$), and 2-km ($z = 3.60$, $p < .001$) distances. The run and walk were equivalent for the 12-min test ($z = 0.80$, $p > .289$). The
- D.

¹ Determined by the method of adding Zs (Rosenthal, 1978).

² Determined by Hedges' Q, (Hedges & Olkin, 1985).

Table 1. Basic Validity Findings

Study	Year	Sample Size	Validity Coefficient	z	SEE
<u>6-min test</u>					
Roul	1998	121	.24	2.66*	4.37
Lipkin	1986	10	.34	.94	2.72
Montgomery	1998	64	.37	2.99*	2.89
Lipkin	1986	10	.54	1.60	2.93
Lipkin	1986	16	.55	2.21*	1.27
Lucas	1999	264	.57	10.46*	4.11
Opasich	2001	311	.59	11.89*	3.55
Faggiano	1997	26	.63	3.56*	3.11
Cahalin	1996	45	.64	4.91*	3.07
Cahalin	1995	30	.67	4.21*	2.75
Zugck	2000	113	.68	8.70*	3.96
Nixon	1996	17	.70	3.25*	3.64
Cahalin	1995	30	.73	4.83*	2.80
Riley	1992	11	.88	3.89*	.
<u>12-min test</u>					
Bernstein	1994	9	.65	1.90*	.
Nakagaichi	1998	25	.73	4.36*	6.42
Nakagaichi	1998	17	.78	3.91*	4.32
<u>1-km test</u>					
Laukkanen	1992	32	.47	2.75*	4.41
Laukkanen	1992	45	.63	4.80*	3.11
<u>1-mi test</u>					
Cureton	1997	92	.27	2.61*	5.39
McCormack	1991	17	.34	1.32	4.33
Jackson	1994	20	.37	1.60	7.34
Cureton	1997	53	.38	2.83*	5.36
McCormack	1991	27	.49	2.63*	3.89
Jackson	1994	21	.55	2.62*	10.27
Draheim	1999	23	.73	4.15*	7.24
Rintala	1992	19	.81	4.51*	5.86
McCormack	1991	15	.82	4.01*	4.96
<u>2-km test</u>					
Laukkanen	1993	44	.31	2.05*	7.32
Laukkanen	1992	32	.49	2.89*	4.36
Laukkanen	1993	32	.52	3.10*	5.12
Oja	1991	35	.58	3.75*	8.06
Laukkanen	1989	79	.61	6.18*	7.53
Laukkanen	1992	45	.72	5.88*	2.78
Laukkanen	1993	35	.73	5.25*	4.78
Oja	1991	29	.74	4.85*	4.51
Laukkanen	1989	80	.75	8.54*	6.28
<u>Miscellaneous tests</u>					
Mercer	1998	14	.83	3.94*	1.80
Singh	1994	19	.88	5.50*	1.95

Note. "Study" = senior author. "SEE" = standard error of estimate.

"." = missing data. * $p < .05$, one-tailed.

Table 2. Summary of Walk Test Validity Results

Test	# ^a	n ^b	Walk r ^c	Z _{null} ^d	Run r ^e	Diff ^f	Z _{diff} ^g	Sig. ^h
<i>Fixed-Time</i>								
6 min	14	1068	.564	17.66	.481	-.083	-3.05	<.004
12 min	3	51	.738	5.87	.789	.051	.80	>.289
<i>Fixed-Distance</i>								
1 km	2	77	.570	5.34	.779	.209	3.33	<.001
1 mile	9	287	.464	8.76	.631	.167	3.88	<.001
2 km	9	411	.635	14.16	.737	.102	3.60	<.001

^aNumber of samples.^bCumulative sample size.^cWeighted average correlation using Fisher's *r*-to-*z* transformation; weights were (*n* - 3) where *n* was the sample size.^dTest of *p* = 0 by the method of adding *z*s (Rosenthal, 1978).^eWeighted average correlation for run tests from Vickers (2001a, 2001b.)^fDifference = (average for run - average for walk).^g*z*-value for run-walk difference with the run average treated as a fixed value (Hays, 1963, pp. 528-532).^hSignificance of run-walk difference.

walk test was superior for the 6-min test (*z* = -3.05, *p* < .004).⁴

- D. Longer walks tended to be more valid than shorter tests. For fixed-time tests, the 12-min walk (*r* = .738) was significantly (*z*_{diff} = 1.95, *p* < .026, one-tailed) more valid than the 6-min walk (*r* = .564). The picture was less certain for fixed-distance tests. If the tests were ordered perfectly by length, the validity of the 1-mile walk would have fallen between the 1-km and 2-km tests. Instead the 1-mile walk had the lowest average validity (*r* = .464). The 2-km walk (*r* = .635) was more valid than the 1-km walk (*r* = .570). However, age contributed to this confusion.⁵ When only adult samples were considered, the 1-mile walk

³ *Z*-scores, including the differences between tests, were computed using Fisher's *r*-to-*z* transformation (Hays, 1963, pp. 528-532).

⁴ *Z*-scores, including the differences between tests, were computed using Fisher's *r*-to-*z* transformation (Hays, 1963, pp. 528-532).

⁵ Appendix A lists the studies from lowest to highest validity coefficients. Younger samples clearly tended to be listed first for the 1-mile walk test. In fact, validity was significantly lower (*z* = -2.60, *p* < .014) for children (*r* = .383) than for adults (*r* = .642). Data from Vickers (2001a, 2001b) showed a similar trend (under 16 years, *r* = .575; over 16 years, *r* = .677) in 24 studies of the 1-mile run. The 1-mile walk was less valid than the 1-mile run for children (run, *r* = .575; walk, *r* = .383; *z*_{Diff} = 3.60, *p* < .0007, one-tailed), but not for adults (run, *r* = .677; walk, *r* = .642; *z*_{Diff} = 0.49, *p* > .353, one-tailed).

validity ($r = .642$) was slightly higher than the 2-km walk validity. The weak general tendency toward higher validity for longer walks was statistically significant ($\chi^2 = 6.01$, 1 df, $p < .015$) when the 6-min and 1-km tests were combined and contrasted with the combined 12-min, 1-mile, and 2-km tests.

- E. If $r = .60$ is a minimum standard for validity (Nunnally & Bernstein, 1994), the 12-min, 1-mile, and 2-km walks were acceptable tests. The 6-min and 1-km walks were below this criterion.

Discussion

Walk tests are valid, but only the 12-min, 1-mile, and 2-km tests met minimum validity standards. Even the average validity of those tests was only borderline acceptable. The 12-min walk test may be an exception to this generalization, but there is too little evidence available at this time to place much confidence in the higher average validity for that test. Note should also be taken of the fact that the average validity for the 6-min and 1-km tests was just below the minimum validity standard. A few additional studies with higher validities could change the inferences for those tests as well. Thus, it should be remembered that the validity difference between the shorter and longer tests achieved statistical significance only when the tests were grouped and analysis was limited to adult samples. The overall data trends were too weak to conclude that there is a sound empirical basis for choosing among the walk tests at this time.

The inference that longer tests are more valid than shorter tests should be viewed with caution, but not discounted all together. This suggestion should be viewed with caution because it was reached in several steps. The results, therefore, might be viewed with skepticism because they involved excessive data manipulation. However, the suggestion that the trend is probably real rests partly on evidence not covered in this review. The validity of fixed-distance run tests increases with the logarithm of distance up to 2 km (Vickers, 2001a, 2001b). If walk tests are analogous to run tests, the fact that validity increases with the logarithm of distance implies that the tests examined here will show only small differences. A weak trend in the expected direction, therefore, may be all that could be expected.

Multivariate Walk Test Equations

Multivariate walk test equations combine walk time (t_w) with other information to improve the precision of VO_{2max} estimates. This section examines two multivariate equations, the Rockport Fitness Walk Test (RFWT; Kline, Porcari, et al., 1987) and the Urho Kaleva Kekkonen Institute Walk Test (UKKWT; Oja, Laukkanen, et al., 1991). Other multivariate walk tests have been developed

(Dolgener, Hensley, et al., 1994; George, Fellingham, et al., 1998), but those tests are not covered here. The sample of participants in the Dolgener, Hensley, et al. (1994) study appears to be atypical. As a result, their equations do not perform well in new samples (Appendix B). The George, Fellingham, et al. (1998) equations have not been studied enough to reach firm conclusions about their value at this time.

Rockport Fitness Walking Test

The RFWT consists of 3 equations developed by Kline, Porcari, et al. (1987). The equations predict VO_{2max} based on the time required to complete a 1-mile walk, heart rate (HR) at the end of the walk, age, weight, and gender.

RFWT Equations. The RFWT equations were developed with data from 88% of 390 volunteers who underwent VO_{2max} testing. The other 12% ($n = 47$) failed to meet established criteria for a valid VO_{2max} test. The participants were divided into two groups. Data from one group ($n = 174$; 92 females, 82 males) were used to develop the equations. Data from the other group ($n = 169$; 86 females, 83 males) were used to cross-validate the equations.

The research design restricted the sample to people between 30 and 69 years of age. Average ages were 46.5 years for males and 48.5 years for females. Average VO_{2max} was 42.2 ($SD = 9.8$) VO_{2max} $ml \cdot kg^{-1} \cdot min^{-1}$ for men and 31.4 ($SD = 8.5$) $ml \cdot kg^{-1} \cdot min^{-1}$ for women. values for study participants were close to what would be expected given the ages of the samples (Fitzgerald, Tanaka, et al., 1997; Wilson & Tanaka, 2000).

Laboratory treadmill measurements of VO_{2max} were the dependent variable in the RFWT equations. Participants ran on the treadmill at a self-selected pace. The test began with the treadmill at 0% grade. The grade was increased 2.5% every 2 min. Participants were encouraged verbally. The test stopped when the individual was unable to continue despite the encouragement. The criteria for determining that a true maximal oxygen uptake had been achieved during the test were (a) VO_2 leveled off during the test despite an increase in work, (b) the respiratory exchange ratio (RER) reached or exceeded 1.10, (c) the exercise HR was less than 15 beats per minute below age-predicted maximal HR. Measured VO_2 uptake was accepted as a valid VO_{2max} when at least 2 of the 3 criteria were met.

Fourteen (14) potential predictors were considered. During each test, participants walked 1 mile as fast as they could. Heart rate was monitored and recorded during the walks. Walk time for the mile and 4 HRs were recorded. The heart rates were the average values for the last 1 min of each one quarter mile of the

Table 3. Rockport Fitness Walk Test Equations

Generalized Equation:

$$VO_{2max} = 132.853 - (.0769*weight) - (0.3877*age) + (6.315*sex) - (3.2649*time) - (.1565*HR)$$

Gender-Specific Male Equation:

$$VO_{2max} = 154.889 - (.0947*weight) - (0.3709*age) - (3.9744*time) - (.1847*HR)$$

Gender-Specific Female Equation:

$$VO_{2max} = 116.579 - (.0585*weight) - (0.3885*age) - (2.7961*time) - (.1109*HR)$$

Note. These equations are taken from Kline, Porcari, et al. (1987). Weight was measured in pounds, age in years, and time in minutes. Sex was coded 0 for females and 1 for males. Heart rate was measured during the last 1 min of the first 1-mile walk.

walk. Each participant completed the walk test at least twice. If the times (t_w s) were within 30 s of each other, the two walks were accepted as providing acceptable performance measures. If the t_w s for the first two tests differed by more than 30 s, "... subsequent walks were performed until this criterion was met" (Kline, Porcari et al., 1987, p. 255). The 14 potential predictors included the 10 walk test measurements plus age, weight, height, and gender. The "best subsets" regression procedure from the BMDP computer package (Dixon, Brown, et al., 1990) was used to establish the final regression equations.

Kline et al. (1987) developed 2 predictive models (Table 3). The first model consisted of a single regression equation for men and women (Generalized Equation). Gender was a predictor in this model. The second model had separate equations for men and women (Gender-Specific Equations). The multiple correlations were high (Generalized, $R = .88$; males, $R = .85$; females, $R = .86$). The standard error of estimate (SEE) was 5.0 for the Generalized Equation, 5.3 for the male equation, and 4.5 for the female equation. If prediction errors were random and normally distributed, true VO_{2max} would have a 95% probability of being within ± 2 SEE of the predicted value.

RFWT Cross-Validations. The RFWT equations have been extensively cross-validated (Table 4).⁶ Each cross-validation

⁶This review covers 10 of 12 studies. Dolgener, Hensley, et al. (1994) were dropped as an outlier (see Appendix B). Ward, Wilkie, et al. (1987) were excluded because they did not indicate which equation(s)

Table 4. Cross-Validation of the RFWT Equations

Study	Year	Gender	n	r	SEE	Bias
Generalized Equation						
Draheim	1999	C	23	.74	7.10	4.75
Coleman	1987	C	90	.79	5.62	0.10
Kittredge	1994	C	25	.81	4.22	10.00
George	1998	C	98	.84	3.58	5.00
Kline	1987	C	169	.88	4.94	-0.10
O'Hanley	1987	C	29	.88	2.71	-5.30
Widrick	1992	C	145	.91	5.10	-0.60
Coleman	1987	F	50	.62	5.49	1.40
George	1998	F	59	.71	2.96	6.00
Fenstermaker	1992	F	16	.78	2.07	-0.15
O'Hanley	1987	F	19	.84	2.28	-6.80
Widrick	1992	F	75	.86	4.34	1.40
Stanforth	1999	F	36	.89	4.10	0.60
Coleman	1987	M	40	.79	5.70	-1.50
George	1998	M	39	.79	3.86	3.60
O'Hanley	1987	M	10	.81	3.17	-2.50
Widrick	1992	M	70	.88	5.18	-2.80
Stanforth	1999	M	31	.89	5.29	-2.20
Gender-Specific Equations						
Zwiren	1991	F	38	.73	4.51	1.50
Fenstermaker	1992	F	16	.79	2.02	0.13
Widrick	1992	F	75	.85	4.48	0.70
Kline	1987	F	86	.86	3.83	-0.10
Stanforth	1999	F	36	.87	4.44	0.50
Kline	1987	M	83	.84	5.70	-0.30
Widrick	1992	M	70	.88	5.18	-2.90
Stanforth	1999	M	31	.89	5.29	-2.60

Note. Results are grouped by equation and gender (C = Combined male and female; F = Female; M = Male). Studies are ordered within groups based from lowest to highest cross-validation coefficient. Multiple correlations in development were $R = .88$ for the Generalized Equation, $R = .86$ for the female Gender-Specific equation, and $R = .85$ for the male Gender-Specific equation.

determined the age, gender, weight, HR, and t_w for new samples of people. The values of these predictors were inserted into the RFWT equations to predict each individual's VO_{2max} . Each cross-validation also included a laboratory measurement of VO_{2max} . Correlation coefficients relating the predicted and measured

they examined. This omission made it impossible to determine where their study fit into the overall body of cross-validation evidence.

Table 5. Summary of RFWT Equation Results

Equation	Gender	$\sum n$	Cross- Validation	SEE	Bias
			r		
Generalized	Combined	579	.87	4.80	1.04
	Female	255	.79	3.92	1.64
	Male	190	.85	4.94	-1.10
Gender-Specific	Female	251	.84	4.10	0.48
	Male	184	.86	5.43	-1.68

Note. Average values for r were computed with the Fisher r -to- z transformation and weighted by $(n - 3)$ where " n " was the sample size, then reversing the r -to- z transformation. Multiple correlations in development were $R = .88$ for the Generalized Equation, $R = .86$ for the female Gender-Specific Equation, and $R = .85$ for the male Gender-Specific Equation.

VO_{2max} values were computed. These correlations, known as cross-validation coefficients, are the focus of this section.

Three aspects of the cross-validations are important. First, equations developed in one sample may be weak predictors of individual differences when applied to data from a new sample. In this case, the average cross-validation coefficients indicated that the predicted VO_{2max} values were strongly related to the observed values. The average coefficients were $r = .87$ for the Generalized Equation, $r = .84$ for the female Gender-Specific Equation, and $r = .86$ for the male Gender-Specific Equation.

Equations also may be biased when applied to data from new samples. Bias occurs when estimated values tend to be consistently lower or consistently higher than observed values of the criterion. The RFWT equations were biased because the average predicted value was $1.04 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ too high for the Generalized Equation, $0.48 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ too high for the female Gender-Specific Equation, and $1.68 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ too low for the male Gender-Specific Equation. The presence of bias was not surprising since statistical considerations associated with using a sample to represent a population make it very likely that at least some bias will be present in any cross-validation. The important point in the present case, therefore, was that the biases were too small to be of practical or theoretical importance.⁷

⁷This interpretation was reached by converting the bias estimates to effect sizes (ESs). The bias was divided by estimates of the standard deviation of VO_{2max} ($SDVO_{2max}$). $SDVO_{2max}$ is $\sim 6.00 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ for samples of people who are similar in age and activity level. $SDVO_{2max}$ increases to $\sim 8.00 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ to $\sim 10.00 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ when wider ranges of age and activity levels are represented (e.g., Kline, Porcari, et al., 1987).

Shrinkage is a third criterion for evaluating the cross-validation performance of multiple regression equations. Regression equations developed using data from one sample typically are less accurate when the equation is applied to data from a new sample. The difference between the original accuracy and the accuracy in the new sample is shrinkage.⁸ The shrinkage of the RFWT equations was trivial, amounting to .01 for the Generalized Equation and .02 for the female Gender-Specific Equation. For males, the average cross-validation coefficient actually was .01 larger than the original multiple correlation.⁹

Table 5 also provides a basis for evaluating the utility of the Gender-Specific Equations. Those equations would be useful if they improved significantly on the Generalized Equation. The cross-validation coefficients for the Generalized Equation in unisex samples are the proper comparison bases? for determining the additional variance explained by the Gender-Specific Equations. These coefficients remove gender differences in performance and VO_{2max} from the analysis.

The Gender-Specific Equations were slightly more accurate than the Generalized Equation. The average cross-validation coefficient for the Gender-Specific Equation for males was $r = .86$. This figure was .01 higher than the average cross-validation coefficient obtained when the Generalized Equation was applied to males. The difference favoring the Gender-Specific Equation was larger (.05) for women ($r = .79$ vs. $r = .84$), but an outlier data point made the trend misleading. Removing Coleman, Wilkie et al. (1987) from the analysis, the average cross-validation r for the Generalized Equation increased to .83, only .01 less than the cross-validation r for the gender-specific equation.

Discussion

The RFWT equations cross-validated well. The average cross-validation coefficient was high ($r > .84$), shrinkage was low ($\leq .02$), and bias was minor ($ES \leq 0.28$). The evidence also provided reason to prefer the Generalized Equation to the Gender-Specific

Pairing the largest bias with the smallest standard deviation yields, $ES \approx 0.28$ (i.e., $1.68/6$). All other combinations yield $ES \leq 0.21$. Cohen's (1988) widely used criteria set $ES \geq 0.20$ as the lower boundary for an effect with practical or theoretical importance (Cohen, 1988).

⁸Paraphrasing Wherry (1984, p. 74) the average cross-validation coefficient will be lower than the original multiple correlation because the initial regression computations fit errors of measurement as well as real trends in the original data. The adjustments to regression coefficients to fit the unique errors of the development sample do not apply to the measurement and sampling errors in a new sample. Thus, less variance will be explained, and the average value of the correlation coefficient will be lower. The lowering is shrinkage.

⁹ Greater accuracy is possible because shrinkage is an average effect, not an inevitable occurrence.

Equations. The Gender-Specific Equations were slightly more accurate in cross-validation, but the gain was too modest to replace a single equation with separate gender equations.¹⁰

UKK Walk Test

The UKKWT research provided an independent replication of key elements of the RFWT findings. The study participants were drawn from a different population. The method of developing the predictive equations was different. The strategy used in cross-validating the equations was different. Nevertheless, the results reinforced key points identified for the RFWT. In addition, the UKKWT studies reported the predictive accuracy of sample-optimized regression equations. This information provided a different frame of reference for interpreting the accuracy of the basic UKKWT equations.

Equation Development. Subjects were recruited from participants in a questionnaire study of health conducted in a city in Finland (Oja, Laukkanen, et al., 1991). The study was design provided a representative sample of 20- to 65-year-old men and women in that city. The UKKWT validation study included VO_{2max} tests for 10 men and 10 women selected at random from each of four age groups (20-25 years, 35-40 years, 50-55 years, and 60-65 years). Complete VO_{2max} and walk test data were obtained from 64 subjects, 29 women (age = 39.1 years, SD = 13.4) and 35 men (age = 41.9 years, SD = 14.0):

VO_{2max} was measured on a treadmill. Testing began with a 5-min walk at 0% grade. Speed was individually chosen between 4.5 and 5.5 km/hr. After 5 min, the treadmill grade was increased to 5%. The 5% grade was maintained for 2 min after which the grade was increased to 7.5%. Grade subsequently was increased 2.5% every 2 min up until a grade of 20% was reached. Once the 20% grade was reached, speed was increased 0.5 km/hr every 2 min. The measured VO_{2max} was accepted as valid if HR was within 15% of age-predicted maximum, RER was at least 1.0, and blood lactate was at least 4.0 mmol/l. Average VO_{2max} was $34.8 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ (SD = $6.7 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$) for women and $43.1 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ (SD = $9.9 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$) for men.

Walking performance was assessed on a flat 500-m stretch of dirt road. Separate walk tests were performed over distances of 1.0, 1.5, and 2.0 km. Preliminary analyses showed that 2-km performance had the strongest relationship to VO_{2max} , so this

¹⁰The principle of parsimony is the basis for preferring the generalized equation. Parsimony focuses on the trade-off between model complexity and model explanatory/predictive power. Taking the number of parameters as an index of model complexity (Popper, 1959), gender-specific equations increased complexity 67% (from 6 to 10 parameters) with only a 1% improvement in accuracy.

distance provided the performance measures for predicting VO_{2max} . Average t_w was 16.9 min ($SD = 1.2$ min) for women and 15.2 min ($SD = 1.4$ min) for men.

The UKKWT equations combined t_w with age, HR, and either height and weight or body mass index (BMI). Walk time was entered first, followed by age, then HR. After these variables were entered, weight and height were added to the equation as separate predictors or as a single BMI (i.e., weight/height²) predictor. Equations were developed separately for women and men. The model with weight as a predictor was slightly more accurate for women. The model with BMI was slightly more accurate for men. The BMI equations were adopted for subsequent studies.

Cross-Validation. The UKKWT research findings are summarized in Table 6. The major inferences from the data are:

- A. The equations were accurate in the development sample. The multiple correlations in those samples were comparable to the values for the corresponding gender-specific RFWT equations (males, UKKWT $R = .83$ vs. RFWT $R = .85$; females, UKKWT $R = .84$ vs. RFWT $R = .86$).
- B. Shrinkage was substantial. UKKWT cross-validation coefficients were substantially lower than the original R s (men, $r = .71$; women, $r = .69$).
- C. Bias was somewhat larger than for the RFWT equations. UKKWT equations consistently underestimated VO_{2max} , with an average bias of -3.87 for men and -1.15 for women.
- D. The bivariate predictor-criterion relationships underlying the equations were stable across samples. The correlations relating VO_{2max} to individual predictors varied across samples, but the differences were no larger than expected by chance (t_w , $\chi^2 = 12.26$, 6 df, $p > .056$; age, $\chi^2 = 6.94$, 6 df, $p > .326$; BMI, ($\chi^2 = 3.52$, 6 df, $p > .741$; HR, $\chi^2 = 3.03$, 6 df, $p > .805$).¹¹
- E. The multivariate approach provided significantly better prediction of the criterion than did the univariate approach based on t_w . Adding age, weight, height, and HR to t_w accounted for significantly more variation in VO_{2max} . The added value of these predictors can be seen by comparing the correlation between t_w and VO_{2max} with the multiple R for each sample. The F -test and significance level given under each multiple R in Table 6 show that the increase in predictive accuracy was statistically significant ($p < .029$) in 6 of 7 samples. The combined trend was highly significant ($p < 10^{-5}$).

¹¹Hedges's Q (Hedges & Olkin, 1985) was used to test for significant differences in the correlation coefficients across samples. The Q values were computed applying the SPSS GLM procedure (SPSS, Inc., Chicago, IL, 1998a, 1998b) to Fisher-transformed correlations with ($n-3$) as the weighting factor.

Table 6. Validity of the UKKWT

	Sample						
	Development		Obese		Moderately Active		Highly Active
	F	M	F	M	F	M	M
N =	29	35	45	32	32	35	44
Age (in years)							
M	39.1	42.9	42.4	41.3	40.6	40.2	44.8
SD	13.4	14.0	8.8	8.8	4.5	4.7	5.6
VO ₂ (in ml•kg ⁻¹ •min ⁻¹)							
M	34.8	43.1	27.2	36.6	36.2	44.4	57.6
SD	6.7	9.9	4.0	5.0	6.0	7.0	7.7
Correlation of VO _{2max} with:							
Age	-.43	-.51	-.35	-.45	.02	-.39	-.23
BMI	-.58	-.51	-.35	-.60	-.34	-.48	-.54
t _w	-.74	-.58	-.72	-.49	-.52	-.73	-.31
HR	.04	.09	.07	-.08	.18	-.03	-.18
Multivariate Equations							
Mult R	.83	.84	.79	.75	.58	.83	.67
F ^a	3.64	12.54	3.75	6.63	.90	5.01	8.32
p<	.028	.001	.019	.002	.457	.007	.001
Cross r			.77	.75	.55	.79	.60
F ^b			.66	.00	.28	1.25	1.26
p>			.652	.999	.922	.311	.301
Bias			-0.9	-4.3	-1.5	-3.3	-4.0
SEE ^c	3.3	5.1	2.55	3.31	5.01	4.29	6.16

Note. Development = Oja, Laukkanen, et al. (1991); obese samples = Laukkanen, Oja, et al. (1992); moderately and highly active samples = Laukkanen, Oja, et al. (1993). M = Male, F = Female. Mult R = multiple correlation coefficient for the sample-specific equation. Cross r = cross-validation coefficient for UKKWT equation. Bias = predicted minus observed score. SEE = standard error of estimate.

^a $F = MS_{reg}/MS_{res} = [(SS_{reg}/df_{reg})/(SS_{res}/df_{res})] = [(R^2 - r_{tw}^2)/3]/[(1 - R^2)/(n - 5)]$. F is the F-test, MS, SS, and df are the mean square, sum of squares, and degrees of freedom respectively. Subscripts "reg" and "res" indicate that the statistic refers to the regression and the residuals, respectively. R² is the squared multiple correlation coefficient, r_{tw}² is the squared correlation of VO_{2max} with t_w, and n is sample size. MS_{reg} has 3 df because the computations reflect variance explained by age, body mass index (BMI), and heart rate (HR).

^b $F = MS_{reg}/MS_{res} = [(SS_{reg}/df_{reg})/(SS_{res}/df_{res})] = [(R^2 - \text{Cross-validation } R^2)/4]/[(1 - R^2)/(n - 5)]$ where n is the sample size.

^cComputed from reported data as $[\sqrt{(1 - R^2)}] \cdot SD$ where SD is the sample standard deviation.

F. The UKKWT equations were nearly optimal in each sample. Each study reported a regression equation developed to optimize the prediction of VO_{2max} in that sample. These sample-optimized equations used the same predictors as the UKKWT equations, but selected regression weights that produced the smallest possible prediction errors for the sample. The multiple Rs for the sample-optimized equations averaged .03 larger (range = .00 to .07) than the cross-validation coefficient. The F-test and significance levels given below the cross-validation Rs in Table 6 show the modest size of these gains. The improvement in predictive accuracy obtained by substituting the sample-optimized equations for the UKKWT equations did not approach significance in any of the 5 samples ($p > .301$ for each).

Discussion. The UKKWT studies underscored the value of a multivariate approach. The predictive utility of this approach was clearly evident. Adding age, BMI, and exercise HR accounted for an average of 22% more of the variance in VO_{2max} than was explained by t_w alone. The cumulative trend was highly significant statistically.

The inclusion of HR in the UKKWT equations may appear problematic. The simple bivariate correlation between this predictor and VO_{2max} is close to zero. The likely explanation is that HR becomes a significant predictor after controlling for the other variables in the equations. The studies did not report the full matrix of correlation coefficients, so this speculation could not be evaluated directly from the data.

The evaluation of shrinkage is more complex. The cross-validation coefficients were substantially smaller than the initial multiple Rs. However, this trend appears to derive from the choice of cross-validation strategies. The UKKWT equations were developed in a sample drawn from a general population. The cross-validation studies were conducted in specialized subgroups from within that general population. As might be expected, VO_{2max} was more variable in the general population than in the subpopulations (Table 6). Other things equal, less variation in the criterion means weaker associations to predictors.¹² As a result, the comparison between the cross-validated equations and the sample-optimized equations is probably a better indicator of shrinkage. The difference in this comparison was only .03, so it is reasonable to conclude that shrinkage was modest after allowing for the restricted variability in VO_{2max} .

Bias was somewhat more problematic for the UKKWT equations than for the RFWT equations. The bias estimate for men was large

¹²This restriction of range effect is a well-known statistical artifact in meta-analyses (Hunter & Schmidt, 1990).

enough to be considered a moderate effect size (Cohen, 1988). The difference for women was too small to be important.

Two important generalizations about multivariate walk test equations can be drawn from the combined RFWT and UKKWT findings. First, multivariate equations are competitive with run tests as aerobic fitness indicators. The maximum validity for run tests is $r \approx .74$ for fixed distance tests and $r \approx .82$ for fixed-time tests (Vickers, 2001a, 2001b). The multiple Rs for the multivariate equations are above this range, but the average cross-validation coefficients fall in this same range. Second, the multivariate character of the equations is important. Considering age, weight, gender, and exercise HR improves the prediction of VO_{2max} .¹³ The only noteworthy problem for the multivariate equations is the possibility that the predicted values have enough bias to limit their utility. The evidence for this problem is limited to the data for the UKKWT applied to men.

Test Precision

Validity coefficients do not provide a complete basis for comparing tests. Validity is a prerequisite for sound testing practices, but focusing solely on validity can be misleading when choosing among valid tests. Test precision should be considered as well. The SEE is the statistical index for test precision. The SEE formula is

$$SEE = \sqrt{(1 - r^2)} * SD.$$

This formula combines test validity (i.e., r) with the sample standard deviation of VO_{2max} (i.e., SD).

The fact that the SEE formula includes SD renders validity an imperfect guide to test precision. Tests with equal validity coefficients could have very different precision. This outcome would result if one test has been validated in samples with large SD s for VO_{2max} (e.g., the general adult population between 30 and 70 years of age) and the other in more homogenous populations (e.g., elite runners).

Table 7 provides SEE estimates for men and women on univariate walk tests and multivariate walk tests.¹⁴ Also, that

¹³The utility of the combined set of predictors has been established. However, some individual predictors may contribute little to the predictive accuracy of the equations. If so, the equations could be simplified by dropping those predictors. This issue is outside the scope of this review.

¹⁴The 6-min walk was not included. All studies of this test involved mixed-gender samples of patients (Appendix A). The sexes presumably were not separated because patient status was more important than gender. The average SEE for the 6-min walk was 3.67.

Table 7. SEE Values for Different Tests

Test	Males	Females
Univariate Walk		
1-mi	6.21	5.89
2-km	6.19	4.05
12-min	5.58	N/A
Average ^a	5.99	4.47
Multivariate Walk		
RFWT General	4.83	3.76
RFWT Specific	5.41	4.01
UKKWT	4.78	3.56
Average ^a	5.01	3.78
Run		
1-mi	4.74	4.71
2-km	5.89	3.45
12-min	3.82	3.35
1.5-mi ^b	4.30	3.90
2-mi ^b	4.70	2.44
3-mi ^b	4.68	2.44
Average ^a	4.69	3.38

Note. "N/A" = not available. Table entries are in $\text{ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$. Define RFWT, UKKWT here.

^aUnweighted average.

^bDistance used in PFT for one branch of military services in the U.S. Department of Defense.

table gives SEEs for run tests covering walk test distances or times. Separate values have been reported for men and women because gender clearly affected test precision. The SEE for males was larger than that for females in all 11 comparisons provided in Table 7.

The 1.5-, 2-, and 3-mile runs have not been considered in previous sections of this paper. These runs were added to Table 7 because they are elements of PFTs in different service branches within the U.S. Department of Defense. These PFTs probably represent the most extensive use of run tests to evaluate aerobic fitness in the adult population. Combining these measures with the 1-mile, 2-km, and 12-min run tests provides a more extensive

¹⁵The 6-min walk was not included. All studies of this test involved mixed-gender samples of patients (Appendix A). The sexes presumably were not separated because patient status was more important than gender. The average SEE for the 6-min walk was 3.67.

basis for comparing walk tests with alternative run tests. The estimated values of SEE for these run tests were derived from data covered in previous reviews of run test validity (Vickers, 2001a, 2001b).

Two general conclusions can be drawn from Table 7. First, multivariate walk tests are more accurate than univariate walk tests ($0.98 \text{ ml}\cdot\text{kg}^{-1}\cdot\text{min}^{-1}$ for men, $0.69 \text{ ml}\cdot\text{kg}^{-1}\cdot\text{min}^{-1}$ for women). Second, run tests are slightly more accurate than multivariate walk tests ($0.32 \text{ ml}\cdot\text{kg}^{-1}\cdot\text{min}^{-1}$ for men and $0.40 \text{ ml}\cdot\text{kg}^{-1}\cdot\text{min}^{-1}$ for women).

The SEEs for the RFWT equations provided another point of interest. The gender-specific SEE was larger than the generalized SEE. The inequality held for both men and women. This finding is further support for the prior suggestion that the Generalized Equation for the RFWT is preferable to the Gender-Specific Equations for that test (p. 11). The previous recommendation was based on a preference for simplicity. The evidence in Table 7 indicates that the preference for simplicity does not entail a loss of accuracy.

The SEE values can be used to choose a field test to estimate $\text{VO}_{2\text{max}}$. Run tests are preferable to multivariate walk tests. Multivariate walk tests, in turn, are preferable to univariate walk tests. This ordering applies if all other things are equal. However, a multivariate walk test might be chosen over a run test if the test will be administered to a population of older individuals who might be at increased risk of injury during the test. A univariate walk test might be preferred to a multivariate walk test to avoid the requirements for collecting and analyzing additional data (i.e., age, weight, exercise HR). The SEE estimates can be used to weigh the gains in terms of reduced risk and ease of administration against the loss of precision in the $\text{VO}_{2\text{max}}$ estimates.

The SEE computations also clearly indicate that choices between tests should not be based solely on validity coefficients. Using the average validity coefficient as the criterion of choice, multivariate walk tests would rank ahead of run tests. Vickers (2001a, 2001b) estimated the upper limit of validity for run tests at $r = .82$. The cross-validation coefficients for multivariate walk tests exceeded this upper limit (cf., Table 4, p. 8). However, the multivariate walk test coefficients were derived in samples with greater variation among subjects than was typical in the studies of run tests. The net result was that multivariate equations explained a larger proportion of the variance in $\text{VO}_{2\text{max}}$, but still left more residual error variance.

Other Issues

Walk tests are valid and competitive with other field measures of aerobic fitness. With these points established, this section briefly considers some other issues that might affect the decision to use a walk test.

Safety concerns can make walk tests an attractive option. These tests merit special attention when test population members are at risk for musculoskeletal injury, heart attacks, and other adverse health consequences from heavier exertion. Properly supervised walk tests are safe even in highly vulnerable populations. Walk tests have been used extensively in severely ill patient populations, primarily those suffering from cardiac disease and chronic lung disease. No significant problems with the walk tests have been reported in the literature on patient populations. Several authors have explicitly mentioned this issue and noted that either no problems or only minor problems arose during testing (Cahalin, Mathier, et al., 1996; Cahalin, Pappagianoulous, et al., 1995; Langenfeld, Mathier, et al., 1990; Nixon, Joswiak, et al., 1996; Riley, McParland, et al., 1992; Roul, Germain, et al., 1998). If the test is safe in these populations, the risk of adverse effects in a healthy, generally active population between 40 and 60 years of age must be minimal.

Practice effects are a concern. People should practice the walk tests to ensure that their performance reflects the best they can do. Several studies have shown that performance improves when a walk test is repeated once or twice. A single practice trial apparently is enough to stabilize performance in healthy normal adults (Jackson & Solomon, 1994).

The fitness of the population being tested may be a concern. Walk tests may not provide sufficient challenge to permit fit, active individuals to utilize their full aerobic capacity (e.g. Widrick, Ward, et al., 1992). If so, walk tests will systematically underestimate aerobic capacity in such individuals.

Conclusions

Walk tests are valid indicators of aerobic capacity. Simple walk tests (i.e., time to cover 1 mile or 2 km or distance covered in 12 min) satisfy minimum standards for estimating VO_{2max} . Multivariate walk test equations that add age, weight, gender, and exercise HR to walk time provide more accurate estimates. The precision of VO_{2max} estimates provided by the multivariate equations is very close to that of endurance runs, including the runs currently used in PFTs.

References

- American Psychological Association. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.
- Barnett, T., & V. Lewis. (1978). Outliers in Statistical Data. NY: Wiley.
- Bernstein, M. L., J. A. Despars, et al. (1994). "Reanalysis of the 12-min walk in patients with chronic obstructive pulmonary disease." Chest **105**: 163-167.
- Cahalin, L. P., M. A. Mathier, et al. (1996). "The six-minute walk test predicts peak oxygen uptake and survival in patients with advanced heart failure." Chest **110**: 325-332.
- Cahalin, L., P. Pappagianopoulos, et al. (1995). "The relationship of the 6-min walk test to maximal oxygen consumption in transplant patients with end-stage lung disease." Chest **108**: 452-459.
- Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ, Erlbaum.
- Coleman, R. J., S. Wilkie, et al. (1987). Validation of 1-mile walk test for estimating $\dot{V}O_{2\max}$ in 20-29 year olds. Medicine and Science in Sports and Exercise, **19**: S29 (abstract).
- Cooper, H., & L. V. Hedges. (1994). The Handbook of Research Synthesis. New York, Russell Sage Foundation: 301-322.
- Cureton, K. J., M. A. Sloniger, et al. (1997). "Metabolic determinants of the age-related improvement in one-mile run/walk performance in youth." Medicine and Science in Sports and Exercise **29**(2): 259-267.
- Dixon, W. J., M. B. Brown, et al. (1990). BMDP Statistical Software: Volume 2. Berkeley, CA: University of California Press.
- Dolgener, F. A., L. D. Hensley, et al. (1994). "Validation of the Rockport fitness walking test in college males and females." Research Quarterly for Exercise and Sport **65**(2): 152-158.
- Draheim, C. C., N. E. Laurie, et al. (1999). "Validity of a modified aerobic fitness test for adults with mental retardation." Medicine and Science in Sports and Exercise **31**(12): 1849-1854.
- Dunn, O. J. (1961). "Multiple comparisons among means." Journal of the American Statistical Association, **56**: 52-64.
- Faggiano, P., A. D'Aloia, et al. (1997). "Assessment of oxygen uptake during the 6-minute walking test in patients with heart failure: preliminary experience with a portable device." American Heart Journal **134**(2): 203-206.
- Fenstermaker, K. L., S. A. Plowman, et al. (1992). "Validation of the Rockport fitness walking test in females 65 years and older." Research Quarterly for Exercise and Sport **63**(3): 322-327.
- Fitzgerald, M. D., H. Tanaka, et al. (1997). "Age-related declines in maximal aerobic capacity in regularly

- exercising vs. sedentary women: a meta-analysis." Journal of Applied Physiology **83**(1): 160-165.
- Fontenot, E. G. (2001). Validity of the one-mile walk-test as a predictor of aerobic capacity. Master's Thesis. Department of Physiology, Colorado State University, Fort Collins, CO.
- George, J., D., G. W. Fellingham, et al. (1998). "A modified version of the Rockport Fitness Walking test for college men and women." Research Quarterly for Exercise and Sport **69**(2): 205-209.
- Hays, W. L. (1963). Statistics for Psychologists. New York, Holt, Rinehart, Winston.
- Hedges, L. V., & I. Olkin (1985). Statistical Methods for Meta-analysis. Orlando, FL, Academic Press.
- Hunter, J. E., & F. L. Schmidt (1990). Methods of Meta-analysis. Newbury Park, Sage Publications.
- Jackson, A., & J. Solomon (1994). "One-mile test: reliability, validity, norms, and criterion-referenced standards for young adults." Medicine, Exercise, Nutrition and Health **3**(6): 317-322.
- Kittredge, J. M., J. H. Rimmer, et al. (1994). "Validation of the Rockport fitness walking test for adults with mental retardation." Medicine and Science in Sports and Exercise **26**(1): 95-102.
- Kline, G. M., J. P. Porcari, et al. (1987). "Estimation of VO_{2max} from a one-mile track walk, gender, age, and body weight." Medicine and Science in Sports and Exercise **19**(3): 253-259.
- Langenfeld, H., B. Schneider, et al. (1990). "The six-minute walk: an adequate exercise test for pacemaker patients?" Pacing Clin Electrophysiol **13**(12 pt 2): 1761-1765.
- Laukkanen, R., P. Oja, et al. (1989). A walking test for assessing the cardiorespiratory fitness of women and men. Jyvaskyla Congress on Movement and Sport in Women's Life, University of Jyvaskyla, Jyvaskyla, Finland.
- Laukkanen, R., P. Oja, et al. (1992). "Validity of a two kilometre walking test for estimating maximal aerobic power in overweight adults." International Journal of Obesity **16**: 263-268.
- Laukkanen, R. M. T., P. Oja, et al. (1993). "Criterion validity of a two-kilometer walking test for predicting the maximal oxygen uptake of moderately to highly active middle-aged adults." Scandinavian Journal of Medical Science in Sports **3**: 267-272.
- Lipkin, D. P., A. J. Scriven, et al. (1986). "Six minute walking test for assessing exercise capacity in chronic heart failure." British Medical Journal **292**: 653-655.
- Lucas, C., L. W. Stevenson, et al. (1999). "The 6-min walk and peak oxygen consumption in advanced heart failure: aerobic capacity and survival." American Heart Journal **138**: 618-624.
- McCormack, W. P., K. J. Cureton, et al. (1991). "Metabolic determinants of 1-mile run/walk performance in children." Medicine and Science in Sports and Exercise **23**(5): 611-617.

- Mercer, T. H., P. F. Naish, et al. (1998). "Development of a walking test for the assessment of functional capacity in non-anaemic maintenance dialysis patients." Nephrology Dialysis Transplantation **13**(8): 2023-2026.
- Montgomery, P. S., & A. W. Gardner (1998). "The clinical utility of a six-minute walk test in peripheral arterial occlusive disease patients." Journal of the American Geriatric Society **46**: 706-711.
- Nakagaichi, M., & K. Tanaka (1998). "Development of a 12-min treadmill walk test at a self-selected pace for the evaluation of cardiorespiratory fitness in adult men." Applied Human Science **17**(6): 281-288.
- Nixon, P. A., J. L. Joswiak, et al. (1996). "A six-minute walk test for assessing exercise tolerance in severely ill children." Journal of Pediatrics **129**: 362-366.
- Nunnally, J. C., & I. H. Bernstein (1994). Psychometric Theory (3rd ed.). New York, McGraw-Hill.
- O'Hanley, S., A. Ward, et al. (1987). Validation of a one-mile walk test in 70-79 year olds. Medicine and Science in Sports and Exercise, **29**: S28 (abstract).
- Oja, P., R. Laukkanen, et al. (1991). "A 2-km walking test for assessing cardiorespiratory fitness of healthy adults." International Journal of Sports Medicine **12**: 356-362.
- Opasich, C., G. D. Pinna, et al. (2001). "Six-minute walking performance in patients with moderate-to-severe heart failure." European Heart Journal **22**: 488-496.
- Popper, K. (1959). The Logic of Scientific Discovery. New York, Basic Books.
- Riley, M., J. McParland, et al. (1992). "Oxygen consumption during corridor walk testing in chronic cardiac failure." European Heart Journal **13**: 789-793.
- Rintala, P., J. M. Dunn, et al. (1992). "Validity of a cardiorespiratory fitness test for men with mental retardation." Medicine and Science in Sports and Exercise **24**(8): 941-945.
- Rosenthal, R. (1978). "Combining results of independent studies." Psychological Bulletin **85**(1): 185-193.
- Roul, G., P. Germain, et al. (1998). "Does the 6-min walk test predict the prognosis in patients with NYHA class II or III chronic heart failure?" American Heart Journal **136**: 449-457.
- Singh, S. J., M. D. L. Morgan, et al. (1994). "Comparison of oxygen uptake during a conventional treadmill test and the shuttle walking test in chronic airflow limitation." European Respiratory Journal **7**: 2016-2020.
- Solway, S., D. Brooks, et al. (2001). "A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain." Chest **119**: 256-270.
- SPSS, Inc. (1998a). *SPSS Advanced Statistics*. Chicago: SPSS, Inc.
- SPSS, Inc. (1998b). *SPSS Base 8.0 Applications Guide*. Chicago: SPSS, Inc.

- Stanforth, P. R., Wilmore, J. H., et al. (1999). Submaximal aerobic fitness evaluation. Final Report. Department of Kinesiology and Health Education, The University of Texas at Austin, Austin, TX.
- Vickers, R. R., Jr. (2001a). "Running performance as an indicator of VO_{2max} : distance effects." Technical Report 01-20. San Diego, CA: Naval Health Research Center.
- Vickers, R. R., Jr. (2001b). "Running performance as an indicator of VO_{2max} : a replication of distance effects." Technical Report 01-24. San Diego, CA: Naval Health Research Center.
- Ward, A., S. Wilkie, et al. (1987). Estimation of VO_{2max} in overweight females. Medicine and Science in Sports and Exercise, **19**: S29 (abstract).
- Widrick, J., A. Ward, et al. (1992). "Treadmill validation of an over-ground walking test to predict peak oxygen consumption." European Journal of Applied Physiology and Occupational Physiology **64**(4): 304-308.
- Wherry, R. J., Sr. (1984). Contributions to Correlational Analysis. Orlando, FL: Academic Press.
- Wilson, T. M. and H. Tanaka (2000). "Meta-analysis of the age-associated decline in maximal aerobic capacity in men: relation to training status." American Journal of Physiology Heart and Circulation Physiology **278**: H829-H834.
- Zugck, C., C. Kruger, et al. (2000). "Is the 6-minute walk test a reliable substitute for peak oxygen uptake in patients with dilated cardiomyopathy?" European Heart Journal **21**(7): 540-549.
- Zwiren, L. D., Freedson, P. S., et al. (1991). Estimation of VO_{2max} : a comparative analysis of five exercise tests. Research Quarterly for Exercise and Sport, **62**(1), 73-78.

Appendix A

Descriptive Summary of Studies Reviewed

Table A-1. Fixed-Distance Walk Tests

Study	Year	n	Age	S	P	M	VO _{2max}		Time		Obs. r	Adj.		z	SEE
							SD	Resid	M	SD		r	r		
<i>1-km</i>															
Laukkanen	1992	32	41.3	1	0	36.6	5.0	-1.1	486	42	.470	.538	2.75	4.41	
Laukkanen	1992	45	42.4	2	0	27.2	4.0	-1.8	540	42	.630	.773	4.80	3.11	
<i>1-mi</i>															
Cureton	1997	92	12.7	1	0	53.1	5.6	4.0	510	138	.270	.288	2.61	5.39	
McCormack	1991	17	10.2	3	0	51.9	4.6	1.8	663	151	.340	.427	1.32	4.33	
Jackson	1994	20	22.8	2	0	39.1	7.9	3.3	852	76	.370	.290	1.60	7.34	
Cureton	1997	53	11.5	2	0	45.8	5.8	6.0	612	144	.380	.391	2.83	5.36	
McCormack	1991	27	7.0	3	0	49.6	4.5	-1.8	679	127	.490	.603	2.63	3.89	
Jackson	1994	21	22.3	1	0	49.5	12.3	4.2	770	80	.550	.306	2.62	10.27	
Draheim	1999	23	21.7	3	0	35.4	10.6	-10.1	938	137	.730	.517	4.15	7.24	
Rintala	1992	19	25.9	1	0	40.0	10.0	-3.8	846	161	.810	.638	4.51	5.86	
McCormack	1991	15	12.5	3	0	51.8	8.7	2.6	499	98	.820	.704	4.01	4.96	
<i>2-km</i>															
Laukkanen	1993	44	44.8	1	0	57.6	7.7	21.3	768	78	.310	.246	2.05	7.32	
Laukkanen	1992	32	41.3	1	0	36.6	5.0	-1.1	1002	78	.490	.559	2.89	4.36	
Laukkanen	1993	32	40.6	2	0	36.2	6.0	6.6	990	72	.520	.520	3.10	5.12	
Oja	1991	35	42.9	1	0	43.1	9.9	6.1	912	84	.580	.396	3.75	8.06	
Laukkanen	1989	79	42.9	1	0	43.3	9.5	6.3	912	84	.610	.437	6.18	7.53	
Laukkanen	1992	45	42.4	2	0	27.2	4.0	-1.8	1136	78	.720	.841	5.88	2.78	
Laukkanen	1993	35	40.2	1	0	44.4	7.0	6.3	918	66	.730	.675	5.25	4.78	
Oja	1991	29	39.1	2	0	34.8	6.7	4.7	1014	72	.740	.702	4.85	4.51	
Laukkanen	1989	80	42.6	2	0	34.5	9.5	5.6	1020	72	.750	.582	8.54	6.28	
<i>Miscellaneous</i>															
Mercer	1998	14	58.8	3	1	17.1	3.2	-13.6	132	33	.830	.940	3.94	1.80	

Note. "Study" gives the senior author. "n" = sample size. "S" is sex (1 = Male, 2 = Female, 3 = Males and females combined). "P" is patient status (0 = No, 1 = Yes). The "M" and "SD" columns indicate the average and standard deviation for the relevant variable. Under VO_{2max} "Resid" = sample mean - predicted mean based on age using equations of Fitzgerald, Tanaka, et al. (1997) and Wilson & Tanaka (2000). "Obs. r" is the observed correlation; "Adj. r" is the correlation adjusted for restriction of range. The z-value is the test of the null hypothesis that $r = 0$ using the Fisher r-to-z transformation ($p < .05$ if $z > 1.95$). "SEE" is the standard error of estimate (see p. 16).

Table A-2. Fixed-Time Walk Tests

Study	Year	n	Age	S	P	M	VO _{2max}		Distance		Obs. r	Adj.		z	SEE
							SD	Resid	M	SD		r			
6-min															
Roul	1998	121	58.5	3	1	17.0	4.5	-13.8	433	108	.240	.313	2.66	4.37	
Lipkin	1986	10	49.0	3	0	32.1	2.9	-2.5	684	25	.340	.600	.94	2.72	
Montgomery	1998	64	68.0	3	1	12.5	3.1	-14.5	355	74	.365	.604	2.99	2.89	
Lipkin	1986	10	53.0	3	1	18.1	3.5	-14.9	559	86	.540	.742	1.60	2.93	
Lipkin	1986	16	57.0	3	1	11.5	1.5	-19.9	402	122	.546	.933	2.21	1.27	
Lucas	1999	264	52.0	3	1	14.0	5.0	-19.4	383	104	.570	.640	10.46	4.11	
Opasich	2001	311	53.0	3	1	14.6	4.4	-18.4	396	92	.590	.706	11.89	3.55	
Faggiano	1997	26	56.0	3	1	15.0	4.0	-16.8	419	120	.630	.773	3.56	3.11	
Cahalín	1996	45	49.0	3	1	12.2	4.0	-22.4	310	100	.640	.781	4.91	3.07	
Cahalín	1995	30	43.5	3	1	9.4	3.7	-27.4	928	410	.670	.826	4.21	2.75	
Zugck	2000	113	54.0	3	1	15.4	5.4	-17.2	466	107	.680	.718	8.70	3.96	
Nixon	1996	17	14.8	.	1	19.0	5.1	-29.3	407	143	.700	.755	3.25	3.64	
Cahalín	1995	30	44.0	3	1	9.6	4.1	-27.0	982	474	.730	.842	4.83	2.80	
Riley	1992	11	65.2	3	1	16.3	N/A	-11.8	374	85	.880	N/A	3.89	N/A	
12-min															
Bernstein	1994	9	67.0	1	1	N/A	N/A	N/A	808	107	.650	N/A	1.90	N/A	
Nakagaichi	1998	25	36.7	1	0	44.9	9.4	5.4	1150	120	.730	.563	4.36	6.42	
Nakagaichi	1998	17	57.0	1	1	28.8	6.9	-2.6	980	90	.780	.735	3.91	4.32	
Miscellaneous															
Singh	1994	19	61.0	3	1	14.2	4.1	-15.6	400	144	.880	.938	5.50	1.95	

Note. The column headed "Study" indicates the senior author. "n" is the sample size for the study. "S" is sex (1 = Male, 2 = Female, 3 = Males and females combined). "P" is patient status (0 = No, 1 = Yes). The "M" and "SD" columns indicate the average value and standard deviation for the relevant variable. "Resid" = sample mean - predicted mean based on age using equations of Fitzgerald, Tanaka, et al. (1997) and Wilson & Tanaka (2000). The "Obs. r" is the observed correlation, while "Adj. r" is that correlation adjusted for restriction of range. The column of z-values is the test of the null hypothesis that $r = 0$ computed using the Fisher r-to-z transformation. The correlation is significant if $z > 1.95$. "SEE" is the standard error of estimate for the sample (see p. 16). An entry of "N/A" in any column indicates that the information was not available for the study.

Table A-3. Summary of Rockport Fitness Walking Test Studies

Sample Characteristics				Cross-Validation Results					
Study	Year	n	Age	M	SD	r	z	SEE	Bias
General Equation									
Combined Samples									
Dolgener	1994	196	19.4	36.6	5.7	.69	11.78	4.15	13.0
Draheim	1999	23	21.7	35.4	10.6	.74	4.25	7.10	4.8
Coleman	1987	90	25.5	49.4	9.2	.79	9.99	5.62	0.1
Kittredge	1994	25	33.3	29.5	7.2	.81	5.29	4.22	10.0
George	1998	98	22.6	42.8	6.6	.84	11.90	3.58	5.0
Kline	1987	169	47.4	37.2	10.4	.88	17.73	4.94	-0.1
O'Hanley	1987	29	73.7	24.8	5.7	.88	7.02	2.71	-5.3
Widrick	1992	145	37.8	42.0	12.3	.91	18.20	5.10	-0.6
Male Samples									
Dolgener	1994	100	19.4	46.3	8.0	.42	4.41	7.27	8.1
Coleman	1987	40	25.5	54.4	9.3	.79	6.52	5.70	-1.5
George	1998	39	24.2	47.8	6.3	.79	6.43	3.86	3.6
O'Hanley	1987	10	73.7	29.4	5.4	.81	2.98	3.17	-2.5
Widrick	1992	70	36.9	49.8	10.9	.88	11.26	5.18	-2.8
Stanforth	1999	31	27.0	51.7	11.6	.89	7.52	5.29	-2.2
Female Samples									
Dolgener	1994	96	19.3	41.2	8.1	.41	4.20	7.38	3.5
Coleman	1987	50	25.5	45.4	7.0	.62	4.97	5.49	1.4
George	1998	59	21.5	39.2	4.2	.71	6.64	2.96	6.0
Fenstermaker	1992	16	69.4	21.1	3.3	.78	3.77	2.07	-0.2
O'Hanley	1987	19	73.7	22.4	4.2	.84	4.88	2.28	-6.8
Widrick	1992	75	38.6	34.8	8.5	.86	10.97	4.34	1.4
Stanforth	1999	36	26.9	40.6	9.0	.89	8.17	4.10	0.6

(continued on next page)

Table A-3. Summary of Rockport Fitness Walking Test Studies

Sample Characteristics				Cross-Validation Results					
Study	Year	n	Age	M	SD	r	z	SEE	Bias
Gender-Specific									
Male Samples									
Dolgener	1994	100	19.4	46.3	8.0	.42	4.41	7.27	8.1
Kline	1987	83	48.4	42.4	10.5	.84	10.92	5.70	-0.3
Widrick	1992	70	36.9	49.8	10.9	.88	11.26	5.18	-2.9
Stanforth	1999	31	27.0	51.7	11.6	.89	7.52	5.29	-2.6
Female Samples									
Dolgener	1994	96	19.3	41.2	8.1	.41	4.20	7.38	3.5
Zwiren	1991	38	33.0	41.3	6.6	.73	5.49	4.51	1.5
Fenstermaker	1992	16	69.4	21.1	3.3	.79	3.86	2.02	0.1
Widrick	1992	75	38.6	34.8	8.5	.85	10.66	4.48	0.7
Kline	1987	86	48.3	32.2	7.5	.86	11.78	3.83	-0.1
Stanforth	1999	36	26.9	40.6	9.0	.87	7.66	4.44	0.5

Note. The column headed "Study" indicates the senior author. "n" is the sample size for the study. The column headed "r" gives the cross-validation coefficient. The column of z-values is the test of the null hypothesis that $r = 0$ computed using the Fisher r-to-z transformation. The correlation is significant if $z > 1.95$. "SEE" is the standard error of estimate for the sample (see p. 16). "Bias" indicates the average difference between the predicted and observed values. A positive bias indicates the prediction is higher than the observed value. A negative bias indicates the prediction is lower than the observed value.

Appendix B
Evaluation of Dolgener, Hensley, et al. (1994) Cross-validation
of RFWT

Additional analyses were carried out to better understand the poor predictive accuracy of the RFWT equations in the Dolgener et al. (1994) data. Was this result the product of special characteristics of the sample or was it a failure of the equations? Two lines of evidence were available to answer this question. First, Dolgener et al. (1994) developed sample specific equations using the same predictors as the RFWT. If the RFWT equations had been at fault, these sample-specific equations would be much more accurate predictors of VO_{2max} than were the RFWT equations. This expectation was not met. The equation for women provided no increase in accuracy at all compared to the corresponding RFWT equation ($r = .41$ for each). The equation for men did improve on the RFWT ($r = .51$ vs. $r = .432$). The improvement was statistically significant ($F_{4,92} = 2.69$, $p < .036$) if this comparison were considered in isolation from the results for females.

The results of individual significance tests must be viewed with caution when multiple tests are performed (Dunn, 1961). When multiple tests are performed, the significance criterion for individual tests should be set higher than if a single test were performed. The increased stringency for individual tests allows for the probability that at least one test will be significant by chance alone (Dunn, 1961). Because two significance tests were performed, an adjusted significance criterion of $p < .025$ (i.e., $.05/2$) was appropriate for the present case. The improvement for males was not statistically significant by this criterion. Thus, the RFWT equations were just as good as the best sample-specific equations. Classifying the equations as having "failed" to cross-validate when they were nearly as accurate as the best possible predictive equation for the sample was not reasonable.

A cross-validation of the Dolgener equations in a sample of females (Fontenot, 2001) provided the second line of evidence indicating that these equations represented an outlier data set. The cross-validation produced a low validity coefficient for the predicted VO_{2max} using either the Generalized Equation ($r = .499$) or the female-specific equation ($r = .458$). Fontenot's (2001) analyses also indicated that the Dolgener et al. (1994) equations had substantial predictive bias. The equations consistently underestimated observed VO_{2max} values.

The combination of weak predictive accuracy for sample-specific equations in the original Dolgener et al. (1994) data with poor cross-validation of the Dolgener equations in a new sample suggests that the Dolgener et al. (1994) sample was atypical. If so, the most important finding in these analyses was

that the RFWT equations fit the Dolgener et al. (1994) data about as well as possible. These points suggest that the Dolgener et al. (1994) study produced outlier values because their sample was atypical. The reasons for this atypical character are not obvious, but the points considered here were sufficient to justify dropping the study as an outlier (Barnett & Lewis, 1978).

REPORT DOCUMENTATION PAGE

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB Control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. Report Date (DD MM YY) 29 July 2002		2. Report Type Technical Interim		3. DATES COVERED (from - to) Feb 2002 - Jun 2002	
4. TITLE AND SUBTITLE Walk Tests as Indicators of Aerobic Capacity				5a. Contract Number: USMC Reimb 5b. Grant Number: 5c. Program Element: 5d. Project Number: 5e. Task Number: 5f. Work Unit Number: 60109	
6. AUTHORS Ross R. Vickers, Jr.				9. PERFORMING ORGANIZATION REPORT NUMBER Report No. 10. Sponsor/Monitor's Acronyms(s) 11. Sponsor/Monitor's Report Number(s)	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Health Research Center P.O. Box 85122 San Diego, CA 92186-5122					
8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES) Chief, Bureau of Medicine and Surgery MED-02 2600 E St NW Washington DC 20372-5300					
12 DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT (maximum 200 words) Military physical fitness tests (PFTs) use distance runs to assess aerobic fitness. Walk tests are alternatives to this practice. This meta-analysis summarized 39 studies (1,927 participants) relating walk test performance to laboratory measures of maximal oxygen uptake (VO_{2max}). The laboratory measures are the accepted gold standard for assessing aerobic fitness. For adults, the average walk test performance- VO_{2max} correlation was $r = .56$ for a 6-min walk, $r = .74$ for a 12-min walk, $r = .57$ for a 1-km walk, $r = .64$ for a 1-mile walk, and $r = .64$ for a 2-km walk. Each average value was highly significant ($p < 10^{-6}$). All of the averages were lower than would be obtained with run tests ($r > .74$), so the review was extended to consider multivariate equations combining walk test performance with age, weight, gender, and exercise heart rate to predict VO_{2max} . These equations have predicted VO_{2max} accurately and cross-validate well. The standard error of estimate (SEE) for VO_{2max} predictions from these equations was only 0.32 to 0.40 $ml \cdot kg^{-1} \cdot min^{-1}$ larger than that for equivalent statistic for run tests. Walk tests are valid and are comparable to run tests as indicators of VO_{2max} when the multivariate approach is used.					
15. SUBJECT TERMS physical fitness, aerobic fitness, walk tests					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UNCL	18. NUMBER OF PAGES 29	19a. NAME OF RESPONSIBLE PERSON Commanding Officer
a. REPORT UNCL	b. ABSTRACT UNCL	c. THIS PAGE UNCL			19b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (619) 553-8429